MedCrave
Step into the World of Research
International Journal of Biosensors & Bioelectronics

Research Article
Open Access
CrossMark

Machine learning models applied to predicate post activity oxygen saturation levels

Abstract

Machine learning is a rapidly growing field with widespread application in various industries, including healthcare. In recent years, significant advancements in machine learning have enhanced our understanding of data analytics and prediction models in analyzing physiological data, such as blood oxygen levels, heart rate, and blood pressure. This paper focuses on physical activity in adolescence during the pandemic period. The study showed that going outside for a short walk can increase blood oxygen levels. Furthermore, the consumption of certain foods can also raise oxygen saturation levels. Maintaining a high blood oxygen level during exercise can improve athletic performance and reduce injury risk. However, during exercise, the blood oxygen level tends to decrease as the heart rate increases to supply more oxygen to the body. When the blood oxygen level drops too low during exercise, it can lead to fatigue, shortness of breath, and even fainting. By tracking heartbeats, blood oxygen levels, and other physiological parameters during exercise with the YAMAY Smart Watch wearable device, individuals can gain insights into their body's response to physical activity and adjust their exercise routine accordingly. After collecting data using wearable devices and mobile apps, machine learning algorithms can be trained to predict changes in physiological parameters of post-activity, providing valuable insights into the effectiveness of different exercises and identifying potential health risks. Our study used supervised machine learning classification algorithms to predict the expected data with the target data. We used K-fold cross-validation techniques to split the dataset into training, validation, and test sets for the supervised machine learning classification algorithms. After testing each of the machine learning models (K-Nearest Neighbor (KNN), Naïve Bayes, and Random Forest), it was found that the Random Forest had the best prediction accuracy of 98.75%. On the other hand, KNN had a poor prediction accuracy, lower than 41.1%. Therefore, the Random Forest model can accurately predict the effect of change on the oxygen saturation level during exercise.

Keywords: machine learning classification algorithms, healthcare, prediction models, physiological parameters, physical activity in adolescents, oxygen saturation level, heartbeats

Volume 8 Issue 1 - 2023

Yu Wang
The New York City College of Technology, USA

Correspondence: Yu Wang, The New York City College of Technology, CUNY, USA Computer Engineering Technology, USA, Tel 7182605893, Email ywang@citytech.cuny.edu

Received: December 30, 2022 | **Published:** March 06, 2023

Introduction

Machine learning is a rapidly growing field with widespread application in various industries, including healthcare. One such application is the development of efficient decision support systems for healthcare. In recent years, significant advancements in machine learning have enhanced our understanding of data analytics and prediction models. An abundant amount of available data would make predictive analysis more accurate. This can be seen in healthcare and medicine, where large amounts of medical records and daily collected data are outlets for algorithms to prove themselves effective and have accurate predictions.[1] There are several subtypes of machine learning: supervised, semi-supervised, unsupervised, and reinforcement learning. Supervised learning utilizes labeled datasets to train algorithms and then classify the data or predict target data. Some examples of algorithms associated with supervised learning include Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbor (KNN), Random Forest, and Decision Tree. Machine learning provides several innovations in healthcare that enhance present treatments and diagnoses. Currently, the focus is on attempting to diagnose diseases and detect abnormalities from data and different imaging scans using biometrics and deep learning.[2] An example is using the KNN algorithm to diagnose heart disease patients more accurately than neural network ensembles.[3] If highly reliable algorithms produce consistent results of high accuracy, it will create automation of a task, thus removing the possibility of human error. This automation will give healthcare workers more time to do tasks requiring more attention or manual work.[4,5] Although the many advancements in machine learning, the goal of automation and the official use of algorithms in healthcare still need to be achieved.

The COVID-19 pandemic has caused disruptions to daily routines and extracurricular activities, limiting opportunities for physical exercise for many adolescents. Our research was initiated with a focus on using machine learning in extracurricular activities, particularly on adolescents' physiological parameters in exercise during the COVID-19 pandemic, such as blood oxygen levels, heart rate, blood pressure, and the time interval of physical activity. The oxygen saturation level (SpO2 level) is a critical physiological parameter indicating the amount of oxygen carried in the blood. For most people, a typical pulse oximeter reading for the SpO2 level is between 95% and 100%. However, various factors can influence this level, including physical activity, medical conditions, and environmental factors. In our research, by using The YAMAY Smart Watch wearable device to collect data from participants (mainly adolescents). By tracking heartbeats, blood oxygen levels, and other physiological parameters before and after exercise, individuals can gain insights into their body's response to physical activity and adjust their exercise routine accordingly. For example, our study showed that going outside for a short walk can increase blood oxygen levels. Additionally, the consumption of certain foods can also raise oxygen saturation levels. However, during exercise, the blood oxygen level tends to decrease as the heart rate increases to supply more oxygen to the body. When the blood oxygen level drops too low during

exercise, it can lead to fatigue, shortness of breath, and even fainting. Therefore, it is crucial to maintain a healthy SpO2 level, particularly during exercise, to ensure that the body receives enough oxygen to function correctly. The collected participant data includes features gathered from surveys such as age, biological sex, body mass index, as well as data monitored through a wearable device such as heart rate, blood pressure, blood oxygen level, and time duration of exercise. The collected data will be pre-processed and imported into an integrated development environment (IDE). After that, the supervised learning algorithms, including Naïve Bayes, KNN, and Random Forest, will be applied to the training dataset. The performance of the algorithms will be analyzed using the testing set to determine which algorithm is most accurate in predicting the occurrence of SpO2 level changes in adolescents. It can provide valuable insights into the effectiveness of different exercises.

The approach and experiment setup

Data acquisition and dataset construction

To ensure accurate and meaningful results, we carefully designed the data collection phase. This involved selecting appropriate physical and digital tools to record, collect, and process data. For this experiment, participants used the YAMAY SmartWatch SW023 (Figure 1) to record raw data related to their physical activity and health indicators such as heart rate, blood pressure, and blood oxygen levels. This data was collected and stored using a shared excel file that includes all members willing to participate in the experiment. Next, the dataset was constructed by organizing and processing the collected data into a format suitable for machine learning analysis. The process involves removing irrelevant data points, normalizing the data, and labeling the data according to relevant categories or outcomes.

Figure 1 YAMAY Smart watch SW023 used in the project.

The PPG sensor can distinguish between pulsatile blood volume, which is related to blood volume changes in the arteries and heartbeat, and non-pulsatile volume, which is related to basic blood volume, respiration, sympathetic nervous system, and thermoregulation.[6] The YAMAY SmartWatch uses a green light (565 nm) for the reflective PPG sensors to measure heart rate and blood pressure and a red light (610 nm) for its transmissive PPG sensor to detect blood oxygen levels.[7] Values for these parameters that are outside the typical range for adolescents, such as resting heart rate below 60 bpm, blood pressure above 120/80 mmHg, or blood oxygen saturation below 95%, may be considered risk factors for the development of cardiovascular disease in adulthood.[8,9] BMI is an important indicator of cardiovascular disease risk and is often used in conjunction with other physiological measures such as heart rate, blood pressure, and blood oxygen levels. However, it is important to note that normal cardiac health can differ between sexes, with blood pressure typically lower in females than in males.[10–12] In addition to heart rate, blood pressure, and blood oxygen levels, the dataset includes other measured variables to provide more comprehensive information about the participants' health and activity patterns. These include age, sex, weight, height, and BMI. Additional information about the types of activities in which the participants engaged and the duration and intensity of those activities can provide important insights into their overall health and fitness. Some variables may or may not have been recorded depending on the participants' circumstances, availability, or willingness. By including these features in the dataset, a machine learning algorithm can be trained to accurately predict changes in SpO2 level during various activities (resting, walking, jogging, cycling, swimming, eating, and other exercises). The trained algorithm could identify patterns in SpO2 level changes during exercise, providing personalized recommendations for improving cardiorespiratory fitness (Figure 2).

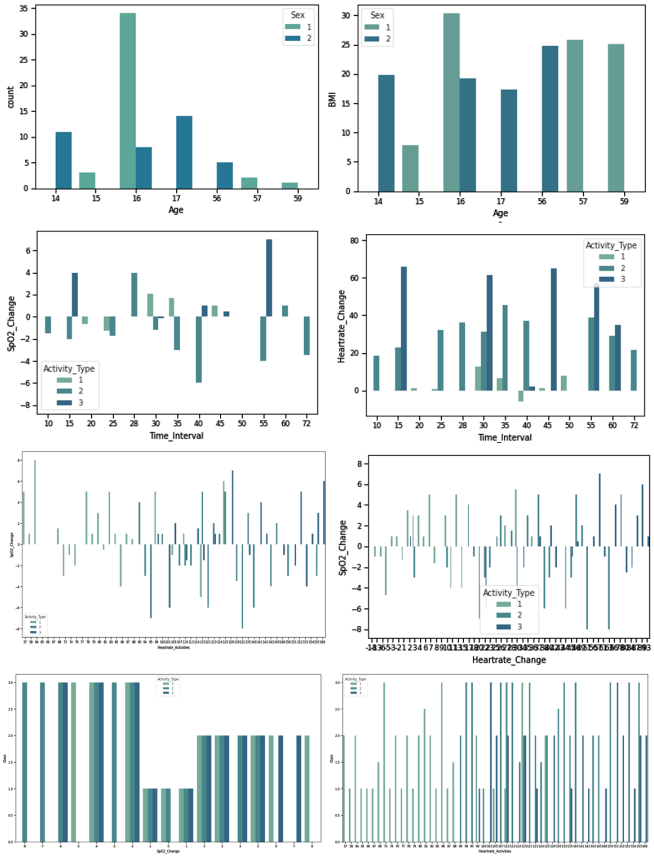

Figure 2 Partial data visualization from our constructed dataset.

The dataset used for the experiment included 78 samples, with 51.3% male and 48.7% female participants. To make data ready for analysis, we eliminated samples that were outside the expected range, reduced the number of features, and selected the relevant spectrum of the data. With feature extraction, we identified the most relevant features in the dataset to build a predictive model. Label encoding allows the algorithms to classify the data effectively and make accurate predictions based on the available features. For example, the activity type in the experiment was assigned a numerical label, with 1 representing eating, 2 representing walking, and 3 representing exercise (jogging, cycling, etc.). Similarly, sex was assigned a numerical label, with 1 representing male and 2 representing female. The target variable for the classification algorithm was determined based on the change in SpO2 levels before and after each activity. The target values can be assigned numerical values using label encoding: class 1 can be assigned to the target value for a change in SpO2 levels within the range of [-1, 1], class 2 can be assigned to the target value for an increase in SpO2 levels, and class 3 can be assigned to the target value for a decrease in SpO2 levels. These labels enable the

Citation: Wang Y. Machine learning models applied to predicate post activity oxygen saturation levels. *Int J Biosen Bioelectron.* 2023;8(1):26–30.
DOI: 10.15406/ijbsbe.2023.08.00228

machine learning algorithms to analyze the data and identify patterns based on the different activities and sex of the participants. With these numerical labels, the algorithms can effectively classify the data and make accurate predictions based on the available features.

From the top leftmost to the bottom rightmost of Figure 2, it shows the number of samples collected for various parameters, including age, sex, BMI, the change of blood oxygen SpO2 level in percentage for each activity, the change of heartbeat rate in beats per minute during each activity, the time duration in minutes for each activity, the change of blood oxygen SpO2 in each target, the resting heart rate distribution for each target before each activity, and the heart rate distribution for each target after each activity. We found that going for a short walk outside can increase the blood oxygen level, and more oxygen is needed to supply to the body. We observed that consuming certain foods can also increase the SpO2 levels in the body, but the heart rate was almost at no change. This can be attributed to the fact that some foods are rich in oxygen-carrying components, which can boost the oxygen levels in the bloodstream. However, when it came to exercising, we noticed that the SpO2 level tended to decrease. This is because the heart rate is increased during exercise. As a result, the SpO2 level dropped as more oxygen was consumed. Additionally, we noticed a change in the participant's heart rates during walking. The body's increased demand for oxygen causes the heart to pump faster to meet the oxygen supply. The collected data also showed that the activity duration length impacted the heart rate and the body's oxygen levels. At high levels of physical activity over a long period, the participant's heart rates increased significantly, and oxygen levels decreased in the blood because the body had to sustain the activity longer.

Applying machine learning models with K-fold cross-validation technique

Supervised machine learning classification algorithms were applied to our constructed dataset with multiple features collected through wearable and mobile devices. The constructed dataset was then split into training and validation sets using the K-fold cross-validation technique to ensure proper model training and evaluation. We employed several popular classification algorithms, including K-Nearest Neighbor (KNN), Naïve Bayes, and Random Forest. Each algorithm was trained and tested on the dataset to predict the expected target. The goal was to predict the expected SpO2 level change in adolescents before and after different events. The accuracy score of each algorithm was studied and compared to identify the algorithm with the highest accuracy in predicting SpO2 level change.

K Neighbors Classifier is a simple algorithm that relies on finding the k-nearest neighbors in the dataset. To find the nearest neighbors, Euclidean distance or Mahalanobis distance is calculated from the input to all known data points. The selection of neighbor k (or n) is based on the size of the dataset. Naïve Bayes is a probabilistic algorithm that assumes independence between features. It is a popular text categorization method that applies Bayes' theorem to separate data based on simple trained features. Essentially, the model assigns labels as feature vectors within a finite set. The main advantage of it is that it only needs a small number of training data sets to begin correctly estimating the parameters necessary for classification. Random Forest is an ensemble algorithm that creates a multitude of decision trees on various sub-samples of the dataset and uses majority voting or averaging to find output. This model combines their predictions for improved accuracy.[13,14] K-fold cross-validation is a method used to validate the accuracy of classification algorithms by dividing the dataset into K subsets, training on K-1 subsets, and testing on the remaining subset. This process is repeated K times, with each subset used as the testing set once. The results from each iteration are then averaged to obtain a final accuracy score. We applied supervised classification algorithms with K-fold cross-validation techniques in the experiment. Table 1 shows the workflow of the experiment of applying these algorithms with K-fold cross-validation, which involves collecting data, preprocessing, feature extraction and selection, data encoding, visualization, machine learning, and model evaluation (Table 1).

Table 1 The workflow of the experiment

Data collection: This step involves collecting data from various sources such as surveys, questionnaires, or measurements. The data should be collected in a structured format to facilitate analysis.
Data preprocessing: This step involves cleaning the data, eliminating data samples that are outside of the expected range, reducing the number of features, and selecting the relevant spectrum of the data.
Feature extraction and selection: This step involves identifying the most relevant features from the dataset that can be used to build the machine learning model.
Data encoding: This step involves transforming categorical data, such as string values, into numerical values that can be used in the machine learning model. For example, a male may be encoded as 1 and a female as 2.
Preprocessing to prepare data for analysis: This step involves setting the value to the desired activity state, such as assigning a numerical value, "1 for eating, 2 for walking, and 3 for exercising".
Setting targets for analysis: This step involves defining the target values that can be used to classify the data based on the degree of change in SpO2 levels before and after the activity (e.g, 1,2,3).
Data visualization: This step involves using various libraries such as pandas, seaborn, matplotlib, numpy, sklearn, etc. to explore the data and identify any patterns or trends.
Machine learning algorithms and multiple models trained: This step involves selecting appropriate machine learning algorithms such as Naïve Bayes, Random Forest, KNN, etc, and training multiple models.
Model evaluation: This step involves evaluating the performance of the models using k-fold cross-validation to determine the best model for the specific problem.
Making predictions and model comparison: This step involves using the Naïve Bayes, Random Forest, and KNN models to make predictions and compare their accuracy.

Results

Machine learning is a complex and iterative process that involves selecting, training, and evaluating models on a given dataset. One of the main challenges in this process is to determine which machine learning algorithm(s) will perform better for a particular task or dataset. Different algorithms have different assumptions, strengths, and weaknesses that can affect their performance depending on the characteristics of the data. We applied three classification algorithms and evaluated their performance using various metrics, such as accuracy score, precision, and standard deviation. The accuracy score of each algorithm was calculated using K-fold cross-validation where K was tuning from 4 to 11. The performance of k-Nearest Neighbor, Random Forest, Naive Bayes with K-folder sizes of 4 and 6 is listed in Table 2. It was found that the Random Forest algorithm had the highest accuracy score of 98.75% with a standard deviation of 2.17%. The Naive Bayes algorithm had a similar accuracy score of 92.30% with a standard deviation of 2.57%, which was also relatively high. However, the k-Nearest Neighbor algorithm performed the worst, with an accuracy score lower than 41.1% among the different hyperparameter tuning since this algorithm is known for its simplicity and ability to handle noisy data, and it may not perform well with this dataset. We also noted that the accuracy score and standard deviation varied with different K-fold sizes of 4 and 6 for each machine learning mode. The performance of the algorithms was influenced by the size of the dataset and the number of folds used in the cross-validation process. Smaller K-fold sizes may lead to higher variance in the estimated performance but can be more computationally efficient (Table 2).

Table 2 Accuracy comparation of supervised machine learning algorithms

	K-folder =4		K-folder =6	
Algorithms	**Accuracy score**	**Standard deviation**	**Accuracy score**	**Standard deviation**
Random Forest	98.75%	2.17%	98.72%	2.87%
Naive Bayes	92.30%	2.57%	92.31%	8.88%
K-Nearest Neighbor (n=26)	28.16%	3.99%	41.03%	7.25%
K-Nearest Neighbor (n=25)	30.72%	5.89%	34.62%	9.68%

The results of this study indicate that the random forest algorithm would be the most appropriate choice for research focusing on changes in SpO2 during exercise in adolescents, especially when using supervised machine learning classification methods. Although there may be modified versions of the algorithms studied that perform better with the given data, the high accuracy score of the random forest algorithm demonstrates its potential for further modification and optimization to achieve even better results in future studies on adolescent cardiac health. Conducting studies on SpO2 changes during exercise in adolescents may benefit from utilizing the random forest algorithm as their primary machine learning classification tool.

Discussion and future work

There are various types of algorithms that may use statistical or mathematical techniques, depending on their code, and each algorithm is suited to a particular area of application. No algorithm is fundamentally better than another. The performance of different algorithms can be affected by various factors, such as the quality of the dataset, the complexity of the problem, and the choice of hyperparameters. The selection of an appropriate algorithm for a given dataset is a challenging task that requires careful consideration of various factors. In addition, to develop a robust algorithm, we recognized the importance of having a large dataset. We have discovered that pulse oximeters are not always accurate and that the actual blood saturation level may vary from what the oximeter reads by 1%~2%. In terms of hardware platforms, the PYNQ FPGA platform can be used in the future. FPGAs will offer several advantages for machine learning applications. It can reduce costs and energy consumption by eliminating the need for specialized hardware. Moreover, FPGAs can provide secure hardware-based solutions for machine learning applications, protecting sensitive data from cyber threats.

Acknowledgments

This work is supported by PSC-CUNY Award # 63382-0051.

Conflicts of interest

There are no conflicting interests declared by the authors.

References

1. Gopinadh S, Abhishek K. Machine learning and big data implementation on health care data. *2020 4th International Conference on Intelligent Computing and Control Systems*. 2020;859–864.
2. Munira F, Jui D, Narayan R C. Machine learning algorithms in healthcare: a literature survey. 2020 *11th International Conference on Computing, Communication and Networking Technologies*. 2020;1–6.
3. Shouman M, Turner T, Stocker R. Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology*. 2012;2(3).
4. Yash V, Shahab T. Evaluation of machine learning architectures in healthcare. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. 2021;1377–1382.
5. Elgendi M, Fletcher R, Liang Y, et al. The use of photoplethysmography for assessing hypertension. *Digit Med.* 2019;2:60.
6. Utami N, Setiawan AW, Zakaria H, et al. Extracting blood flow parameters from photoplethysmograph signals: a review. *The 3rd International conference on instrumentation communications information technology and Biomedical Engineering*. 2013;403–407.
7. Cui W, Ostrander LE. Lee BY. *In vivo* reflectance of blood and tissue as a function of light wavelength. *IEEE Trans Biomed Eng*. 1990;37:632–639.
8. Riley M, Bluhm B. High blood pressure in children and adolescents. *Am Fam Physician.* 2012;85(7):693–700.
9. Kobayashi M, Fukuda S, Takano K, et al. Can a pulse oxygen saturation of 95% to 96% help predict further vital sign destabilization in school-aged children?. *Medicine*. 2018;97(25):11135.
10. Syme C, Abrahamowicz M, Leonard GT, et al. Sex Differences in blood pressure and its relationship to body composition and metabolism in adolescence. *Arch Pediatr Adolesc Med.* 2009;163(9):818–825.

11. Katarya R, P Srinivas. Predicting heart disease at early stages using machine learning: a survey. *2020 International Conference on Electronics and Sustainable Communication Systems*. 2020;302–305.

12. Deng E, Lee E, Shameti D, et al. Applying supervised machine learning algorithms to detect cardiac events. *Fall ASEE Middle Atlantic Section Meeting*. 2021.

13. Hosseini MP, Hosseini A, Ahi K. A review on machine learning for EEG Signal processing in bioengineering. *IEEE Reviews in Biomedical Engineering.* 2021;14:204–218.

14. Abdullah A. Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked.* 2022;30.

Citation: Wang Y. Machine learning models applied to predicate post activity oxygen saturation levels. *Int J Biosen Bioelectron.* 2023;8(1):26–30.
DOI: 10.15406/ijbsbe.2023.08.00228